

Saying and Doing: Does Combining Public Messages with Actions Improve Commitment-Violation Detection by a Frozen LLM Monitor?

Author. CHEN Xuanhui, The University of Hong Kong

Sprint track. Research question 2.2, "Does combining what agents say with what they do improve detection?"

Abstract

Monitors for multi-agent systems usually read either what agents say or what they do. We ask whether one frozen LLM monitor detects broken public commitments more reliably when it sees both. We use the Diners dilemma from the LLM-Trust-Breaking harness (Shi et al.), where each agent writes a private plan, makes a public announcement, and orders a dish. The monitor receives four views of the same trajectory: messages only, actions only, raw messages plus actions, and a structured report that sets quote-checked public commitments next to the observed dish. We froze the monitor prompt and a single threshold (score ≥ 0.6) before evaluation and report accuracy and false-positive rate (FPR) for every view. Three experiments were run with GPT-5.6 Terra, Claude Sonnet 5, and Gemini 3.7 Flash in the player, extractor, and monitor roles. In a balanced five-player, five-round game (27 role combinations, 432 monitor calls), the single-side views stay at 0.5 and so reach only the 50% base rate. Raw combined accuracy is 91.7 to 100% with FPR at most 5.6%, and the structured report reaches 100% accuracy with 0% FPR for all three monitors. A balanced mixed-model table (54 combinations, 864 calls) reproduces this ordering. Natural, unbalanced play (216 calls) agrees, but it has only one independent non-violation. Detection here needs both sides of the record. The results do not measure spontaneous deception or collusion.

1. Introduction

When language-model agents interact, a monitor can see two kinds of evidence. It can see what the agents *say* to each other and what they *do*. A monitor that reads only messages cannot check whether a stated plan was carried out. A monitor that reads only actions cannot know what was promised. Research question 2.2 asks whether combining the two improves detection, compared with either side alone.

We study the simplest version of this question that can be measured cleanly. The behavior to detect is a broken public commitment. A target agent publicly states which dish it will order in a round, and its actual order differs. This is not collusion. It is the observable building block that many collusion monitors rely on: a mismatch between claims and outcomes.

The design holds everything fixed except the evidence the monitor receives. One monitor model, one system prompt, and one decision threshold are frozen before any evaluation. The same trajectories are then shown to the monitor in four views. In the two balanced experiments, every positive trajectory has a matched negative that shares the other players' speech and actions. This makes accuracy and FPR directly comparable across views.

Contributions (new work in this sprint).

1. A four-view monitoring protocol (messages only, actions only, raw combined, structured claim-versus-outcome report) with the prompt, threshold, and data frozen and hashed before scoring.
2. A balanced 2x2 crossing of the target's final public statement with the target's final dish, embedded in a five-player, five-round game whose first four rounds are a shared history.

3. A mixed-model (heterogeneous) variant in which the monitored player sits in a minority seat played by a different model from the majority.
4. A natural-play comparison with no assigned actions, where all orders are the models' own.
5. Full player x extractor x monitor role matrices over three frontier models, with scores, call logs, and costs saved.

2. Related Work

Commitment breaking in repeated games. Shi et al. [1] place LLM agents in repeated n-player games with a three-stage protocol that separates private intent, public announcement, and final action. They measure whether announced commitments are kept and whether breaks were already planned in private. We reuse their Diners game and three-stage turn (the `llm-trust-breaking` harness). Their question concerns agent behavior. Ours concerns the monitor's information: given a fixed set of trajectories, which evidence lets a monitor recover the commitment-versus-action relation?

Information access for monitors. Arike et al. [2] find that LLM monitors can perform *better* with less information (a "less-is-more" effect). They introduce extract-and-evaluate (EaE) monitoring, where one model selects relevant excerpts and another scores them. Our structured report is an EaE-style adaptation with one deliberate difference. The extractor sees only public dialogue, never actions, so it cannot infer commitments backwards from outcomes. Its quotes are checked to be exact substrings of the dialogue before a rule-based tracker adds the observed dish. Unlike their setting, our messages-only view is public speech, not chain-of-thought.

Action-trace consistency. TraceGuard [3] scores agent actions on five dimensions, one of which is action-trace consistency. It separates an untrusted actor from a trusted monitor. TraceGuard is a single-agent control protocol, not a collusion detector. We reuse only its consistency question: does the target's stated action match the action taken in the same round? We did not run the TraceGuard Docker stack or its five-dimension aggregation.

Collusion in multi-agent systems. SCHEME [4] studies coordinated sabotage that no proper subset of agents can complete alone. COLOSSEUM [5] audits collusion in cooperative multi-agent systems and compares action-based regret with communication-based evidence. Both motivate monitors that read speech and action together. Neither is what our monitor is asked to detect.

3. Methodology

3.1 Game

We use the Diners game from `llm-trust-breaking/src/games/diners.py`. Each player orders CHEAP (enjoyment 5, cost 2) or EXPENSIVE (enjoyment 10, cost 8). The bill is split equally, so $\text{payoff} = \text{own enjoyment} - \text{total bill} / N$. Both dishes are legal. Ordering EXPENSIVE is *not* by itself a violation. With N players and the other players' orders held fixed, switching from CHEAP to EXPENSIVE raises one's own payoff by $5 - 6/N$, which is 3.8 for $N = 5$. This creates the social-dilemma temptation. Each round has three stages: a private plan, a public announcement, and an order.

3.2 Label

The label is binary per trajectory and concerns one designated target player. A trajectory is a **violation** if, in at least one round, the target's explicit, unconditional public own-action commitment for that round differs from the target's dish in the same round. Other players' statements are not the target's commitments. A "this round" commitment expires after that round. Tentative preferences and conditional statements are not unconditional commitments.

Reference labels are the *agreement* of two independent reviewer models, GPT-4.1 and Claude Sonnet 4.5. Both read the public dialogue and the actual actions, and each must cite an exact quote. Reviewers never see the extractor output, the monitor scores, or the private plans. Disagreements were not forced to a label. The reviewers differ from every player, extractor, and monitor model.

3.3 Four views

All four views come from the same trajectory and go to the same monitor with the same system prompt.

View	What the monitor receives
Messages only	Public announcements of all players. No dishes, no payoffs.
Actions only	Dishes, payoffs, and round outcome. No announcements.
Raw combined	Public announcements together with dishes, payoffs, and outcome.
Structured report	Target's public commitment events extracted from dialogue (quotes verified), commitment state per round, plus the target's dish and payoff.

Extractor. An LLM extractor reads only the public dialogue up to the current round. It never sees actions, outcomes, labels, private plans, or future rounds. It returns typed events (commitment, revision, retraction, discussion, conditional, ambiguous), each with an exact evidence quote. Events are rejected if the quote is not a contiguous substring of the referenced message, or if speaker, round, or supersession fields are inconsistent. A deterministic tracker then resolves the commitment state per round (policy `latest_explicit`: explicit pre-action revisions are allowed, mere discussion never revises a promise). Only after that is the target's dish appended.

Monitor prompt (frozen; abridged). "You are a monitor of public commitment/action consistency in a Diners game. Estimate the probability that the TARGET player violated at least one explicit, unconditional own-action commitment ... A violation means the player's actual dish differs from their public commitment for the SAME round. CHEAP and EXPENSIVE are both legal ... Do not infer collusion, hidden intent, or private plans ... Messages alone do not establish actual actions; actions alone do not establish commitments. When one side is missing, express uncertainty (probability near 0.5), not invented evidence ... Return ONLY JSON {score, reason}." The structured-report addendum states that the report contains fallible LLM-extracted evidence and that the monitor should return 0.5 when the resolution is unresolved. The full prompts are in `weekend_2_2/experiment.py` (PROMPT), `paired_crossing/run.py` (EVALUATOR_PROMPT), `paired_crossing/commitments.py` (EXTRACTOR_PROMPT), and `paired_crossing/direct_labels.py` (PROMPT). Each run also saves them in `protocol.json`.

3.4 Decision rule and metrics (frozen before evaluation)

The monitor returns a score in $[0, 1]$. The prediction is "violation" if and only if $\text{score} \geq 0.6$. This threshold was set before any LLM monitor run and was never changed. The primary metrics use **all** scored rows. A score of exactly 0.5 therefore counts as a predicted non-violation:

- $\text{Accuracy} = (\text{TP} + \text{TN}) / N$

- $FPR = FP / (FP + TN)$, computed on non-violation rows
- $Recall = TP / (TP + FN)$, reported so that a constant "no violation" prediction is not mistaken for a good detector

As secondary analyses we report coverage (the share of rows with score not equal to 0.5), per-class score means with SD, standard errors clustered by independent source game, the violation-minus-non-violation score gap, AUROC, and Brier score.

3.5 Models and inference settings

Players, extractors, and monitors are drawn from GPT-5.6 Terra (openai/gpt-5.6-terra), Claude Sonnet 5 (anthropic/claude-sonnet-5), and Gemini 3.7 Flash (google/gemini-3.7-flash). All calls go through OpenRouter with the provider pinned and fallback disabled. Each call has a 4096-token output budget and reasoning effort "low". Each monitor request is an independent context, and monitor responses are never cached or reused. The run seed is 20261026. Temperature and seed are sent only where the endpoint supports them: seed for Terra and Gemini, temperature for Gemini only. Remote inference is therefore not guaranteed to be bit-for-bit reproducible. We save every HTTP call, its parameters, the returned model identifier, and the server-reported cost. The protocol, data, views, and source files are SHA-256 hashed before scoring and re-verified at scoring time.

3.6 Experiments

E1. Balanced multi-round game (endogenous_rounds_results/role_matrix_27_20261026; code in endogenous_rounds/ and paired_crossing/run_endogenous_matrix.py). This is the endogenous_rounds process embedded in the paired_crossing 2x2 design. Five diners play five rounds, and the target speaks last. Rounds 1 to 4 form a shared history. The target is constrained to order what it announced (an "honor control"), and the other four players plan, speak, and order from the dialogue they can see. In round 5, the other players speak and lock their orders first. The target's final statement is then one of two commitments (CHEAP or EXPENSIVE), and the target's dish is set to one of two values. The four combinations give two violations and two non-violations with identical surrounding evidence. Each of the three player models generates one new five-player game. Crossing 3 extractors x 3 monitors gives 27 role combinations, 108 stored trajectories (54 positive, 54 negative), and 432 monitor calls. Only 3 source games are independent. Server-reported cost: USD 2.83.

E2. Balanced mixed-model table (minority_player_results; code in minority_players/ and paired_crossing/run_minority_matrix.py). This uses the same five-round crossing as E1, but the five seats are split into three majority seats (Blair, Casey, Drew) and two minority seats (Alex, Evan). The two minority seats share one model, which differs from the majority model. The monitored target is Evan, a minority seat. All six ordered (majority, minority) pairs of the three models are used, giving six source games. Crossing with 3 extractors x 3 monitors gives 54 combinations, 216 trajectories (108 positive, 108 negative), and 864 monitor calls. The monitor is not told which model played which seat. There is no side payment, faction payoff, or private channel. Server-reported cost: USD 6.05.

E3. Natural play (naive_result; code in paired_crossing/run_naive_matrix.py). This is the original paired_crossing three-stage pipeline with *no* crossing. Three diners play two independent rounds per game. Every player plans, speaks, and then orders after reading all announcements, and every order is the model's own. There are no assigned dishes, no statement flips, no honor control, and no rejection or resampling for balance. Each player model plays two games, giving six independent source trajectories. The reviewers agreed

on 5 violations and 1 non-violation. Crossing 3 extractors x 3 monitors reuses these six games and yields 216 monitor calls. Server-reported cost: USD 0.93.

3.7 What counts as a replicate

A source game is one generated table. Role combinations that reuse its speech and actions are repeated measurements, not new samples. Standard errors are clustered by source game. With only 3 (E1) or 6 (E2, E3) games, they are descriptive rather than inferential.

3.8 Controls

- **Positive control (pipeline).** The rule-based monitor `commitment-consistency-rule-v1` (`runs/rule_baseline`) returns 0.5 when one side is missing and otherwise compares commitments with dishes exactly. It was run on 40 scripted trajectories from 10 balanced 2x2 groups (3 players, 3 rounds) with the same views and threshold. It shows that the views, labels, and metrics behave as intended when the comparison is done perfectly.
- **Negative controls (ablations).** In the balanced designs, messages-only and actions-only views are exactly class-balanced *within each view*. Every message transcript appears once as a violation and once as a non-violation, and the same holds for every action record. These views therefore carry no label information. A correct monitor should stay near 0.5 on them, and any accuracy above 50% would indicate leakage.

4. Results

4.1 Positive control

The rule-based monitor reaches 50% accuracy, 0% FPR, and 0% recall on both single-side views (20 TN and 20 FN each). On both two-sided views it reaches 100% accuracy and 0% FPR (20 TP and 20 TN). The pipeline behaves as designed, and it sets the ceiling the LLM monitors are compared against.

4.2 E1: Balanced multi-round game

Each monitor contributes 36 scores per view (18 violations, 18 non-violations) from the same three source games.

Monitor	Messages only Acc / FPR	Actions only Acc / FPR	Raw combined Acc / FPR	Structured Acc / FPR
GPT-5.6 Terra	50.0% / 0.0%	50.0% / 0.0%	91.7% / 5.6%	100.0% / 0.0%
Claude Sonnet 5	50.0% / 0.0%	50.0% / 0.0%	97.2% / 0.0%	100.0% / 0.0%
Gemini 3.7 Flash	50.0% / 0.0%	50.0% / 0.0%	100.0% / 0.0%	100.0% / 0.0%

Table 1. E1 primary metrics at the frozen threshold ($n = 36$ per cell).

All three monitors returned exactly 0.5 on every messages-only and actions-only row, so coverage is 0%. The frozen rule therefore predicts no violation: accuracy equals the 50% base rate, FPR is 0%, and recall is 0%. This is the negative control behaving correctly, not a working detector. On raw combined input, recall is 88.9% (Terra), 94.4% (Sonnet 5), and 100% (Gemini). Terra's one false positive and its two misses explain its lower raw accuracy.

Monitor	View	Mean score, viol.	Mean score, non-viol.	Gap (clustered SE)	AUROC	Brier
GPT-5.6 Terra	Raw combined	0.884	0.055	0.829 (0.097)	0.929	0.083
GPT-5.6 Terra	Structured	0.998	0.000	0.998 (0.001)	1.000	0.000
Claude Sonnet 5	Raw combined	0.900	0.036	0.864 (0.062)	0.983	0.027
Claude Sonnet 5	Structured	0.965	0.000	0.965 (0.002)	1.000	0.001
Gemini 3.7 Flash	Raw combined	1.000	0.000	1.000 (0.000)	1.000	0.000
Gemini 3.7 Flash	Structured	1.000	0.000	1.000 (0.000)	1.000	0.000

Table 2. *E1 continuous-score analysis. Single-side views have mean 0.5 for both classes, gap 0, AUROC 0.5, and Brier 0.25 for every monitor.*

Across the 27 role combinations, raw combined accuracy per combination is either 75% or 100%. Four combinations reach 75%, all with Terra or Sonnet 5 as monitor. Structured accuracy is 100% with 0% FPR in all 27. Gemini's zero standard errors mean that no variation was observed across three games. They are not a population estimate.

4.3 E2: Balanced mixed-model table

Each monitor contributes 72 scores per view (36 violations, 36 non-violations) from six source games.

Monitor	Messages only Acc / FPR	Actions only Acc / FPR	Raw combined Acc / FPR	Structured Acc / FPR
GPT-5.6 Terra	50.0% / 0.0%	50.0% / 0.0%	87.5% / 8.3%	100.0% / 0.0%
Claude Sonnet 5	50.0% / 0.0%	50.0% / 0.0%	97.2% / 0.0%	100.0% / 0.0%
Gemini 3.7 Flash	50.0% / 0.0%	50.0% / 0.0%	100.0% / 0.0%	100.0% / 0.0%

Table 3. *E2 primary metrics at the frozen threshold ($n = 72$ per cell).*

The single-side views again sit at 0.5. The one exception is Terra on actions only: a few rows moved off 0.5 (coverage 2.8%, mean 0.486 in both classes), and none crossed the threshold. Raw combined recall is 83.3% for Terra (6 misses, 3 false positives), 94.4% for Sonnet 5, and 100% for Gemini. Sonnet 5 abstained on one raw row (coverage 98.6%). On the structured report, recall is 100% and FPR is 0% for every monitor. The structured violation-minus-non-violation gaps are 0.999 (Terra, SE 0.001), 0.938 (Sonnet 5, SE 0.002), and 1.000 (Gemini). The corresponding raw combined gaps are 0.749 (SE 0.141), 0.886 (SE 0.025), and 1.000. Raw combined AUROC is 0.873, 0.995, and 1.000, while structured AUROC is 1.000 for all three monitors.

Across the 54 combinations, structured accuracy is 100% in every one. Raw combined accuracy is 50%, 75%, or 100%. Both 50% cases have Terra as monitor and Terra as majority player (a Sonnet 5 minority target, with Sonnet 5 or Gemini as extractor). Because seat pair and source game are confounded, this should not be read as a ranking of minority models. Only the minority seat is monitored, so the design does not estimate an effect of being in the minority.

4.4 E3: Natural play

The six independent natural games contain 5 violations and 1 non-violation. After reuse across extractors, each monitor has 18 rows per view, made up of 15 violation rows and 3 non-violation rows. The 3 non-violation rows are the same game seen through three extractors.

Monitor	Messages only Acc / FPR	Actions only Acc / FPR	Raw combined Acc / FPR	Structured Acc / FPR
GPT-5.6 Terra	16.7% / 0.0%	16.7% / 0.0%	100.0% / 0.0%	100.0% / 0.0%

Monitor	Messages only Acc / FPR	Actions only Acc / FPR	Raw combined Acc / FPR	Structured Acc / FPR
Claude Sonnet 5	16.7% / 0.0%	16.7% / 0.0%	100.0% / 0.0%	100.0% / 0.0%
Gemini 3.7 Flash	16.7% / 0.0%	16.7% / 0.0%	100.0% / 0.0%	100.0% / 0.0%

Table 4. E3 primary metrics ($n = 18$ per cell, only 1 independent non-violation).

Single-side scores are again 0.5 everywhere, so the frozen rule predicts no violation. Accuracy is then 3/18, the share of non-violation rows, and recall is 0%. On the two-sided views, violation means range from 0.944 to 1.000 (clustered SE at most 0.009), and the single non-violation scores exactly 0 for every monitor. The paired structured-minus-raw score difference on violations is small (+0.007, +0.004, and -0.003 for Terra, Sonnet 5, and Gemini). This table shows that the two-sided views track commitment breaks that the models produced on their own in these six games. Its FPR rests on one independent non-violation, so it is not a calibrated FPR.

4.5 Development check

A pipeline check at the root of `endogenous_rounds_results` used one five-player game with `openai/gpt-4o-mini-2024-07-18` as player and extractor and `openai/gpt-4.1-mini` as monitor (16 monitor calls, about USD 0.20). Accuracy was 50%, 50%, 75%, and 100% across the four views, with 0% FPR on every view. Unlike the frontier monitors, `gpt-4.1-mini` did not always abstain on messages only. It gave scores of 0 to both violations, which yields an AUROC of 0.25. The reviewer pair was changed during preparation, before monitoring, so this run is not pooled with E1 to E3.

4.6 Cost

Server-reported costs were USD 2.83 (E1), 6.05 (E2), 0.93 (E3), and 0.20 (development), about USD 10 in total. These figures include preparation calls and retries that returned usage data.

5. Discussion

Main finding. Under a frozen rule, combining speech and action is necessary for this detector, and structuring the combination helps further. With either side missing, all three frontier monitors followed the instruction to stay at 0.5, and the frozen threshold then missed every violation. With both sides present, accuracy rose from the 50% base rate to 87.5 to 100%, and FPR stayed at or below 8.3% on balanced data. The structured claim-versus-outcome report removed every remaining raw-view error in E1 and E2, giving 100% accuracy and 0% FPR in all 81 balanced role combinations.

Why the structured view helps. The raw-view errors come mainly from GPT-5.6 Terra and, less often, Claude Sonnet 5. We did not analyze each error's stated reason. A plausible explanation is that the monitor must find the target's own commitment inside a long five-player, five-round transcript that also contains other players' claims. The structured report does this localization in advance and places the quote next to the dish. This agrees with the extract-and-evaluate result of Arike et al. [2] that filtered evidence can help a monitor. Here the effect is a ceiling: the remaining gap between raw and structured is small, and Gemini 3.7 Flash had no gap at all.

Important caveat: the structured view is close to the label. The label is defined as the relation between the public commitment and the dish, and the structured report puts exactly that relation in front of the monitor. Perfect structured accuracy shows that the extractor kept a usable public quote and that the monitor compared it correctly. It does not show that the monitor discovered hidden intent. The value of the structured view in harder settings (ambiguous, conditional, or revised commitments; deceptive phrasing aimed at the extractor) is untested.

Constructed versus natural behavior. In E1 and E2 the final-round commitments were controlled. In all nine source games, sampling the target freely gave the same dish five times in a row and triggered a patience limit, after which the two statements were requested explicitly. The mismatched dish is assigned by the experiment. These datasets measure the monitor's ability to *read* a mismatch, not how often models create one. E3 is the only experiment with natural actions, and it lacks a balanced negative class.

Heterogeneous groups are not collusion. Mixed-model tables change who speaks, but no agent was told it had a partner and no private channel existed. Differences across the six seat pairs are confounded with the six source games.

Future work. (i) Collect enough natural games under the same frozen threshold to estimate FPR on spontaneous non-violations. (ii) Add ambiguous, conditional, and revised commitments, and adversarial phrasing aimed at the extractor. (iii) Regenerate post-fork history separately for each branch instead of reusing it. (iv) Add a private channel or coordination incentive while holding the public trajectory fixed, so that the question remains about monitor information rather than changes in play. (v) Run TraceGuard's full multi-dimension rubric as an alternative monitor.

References

- [1] Shi, J., Zhang, T. J., Scholkopf, B., Conitzer, V., and Jin, Z. When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. arXiv:2607.05132, 2026. Code: LLM-Trust-Breaking (llm-trust-breaking/).
- [2] Arike, R., Mehta Moreno, R., Subramani, R., Biswas, S., and Ward, F. R. How Does Information Access Affect LLM Monitors' Ability to Detect Sabotage? arXiv:2601.21112, 2026.
- [3] Nguyen, K. L., Nghiem, H., and Tran, T. TraceGuard: Structured Multi-Dimensional Monitoring as a Collusion-Resistant Control Protocol. arXiv:2604.03968, 2026. With Apart Research. Code: trace-guard/.
- [4] Radev, N., Haas, L., Arnav, B., and Bernabeu-Perez, P. The Best-Laid SCHEMES: Coordinated Sabotage and Monitoring in Multi-Agent Systems. arXiv:2605.29178, 2026.
- [5] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E. COLOSSEUM: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198, 2026.

Appendix A. Limitations and Dual-Use / Ethical Considerations

What the results establish. For these three monitor models, this prompt, and the threshold 0.6, a monitor that sees both the target's public commitment and the target's dish detects commitment-action mismatches far more accurately than a monitor that sees only one of them. In this Diners setting, a checked structured report was the most accurate view in every balanced run.

What the results do not establish.

- They do not show that models deceive, collude, or exploit each other. In E1 and E2 the mismatched dish is assigned by the experiment, and the final statements were requested explicitly after a patience limit.
- They do not estimate natural violation rates or a calibrated FPR. E3 has only one independent non-violation.

- They do not rank models. There are only 3 (E1) or 6 (E2, E3) independent source games, and role combinations reuse the same evidence. Zero variance (for example, Gemini) means "not observed in this sample".
- Labels are agreement between two LLM reviewers with quote checks. They were not verified by hand on every transcript.
- Scores are not calibrated probabilities. The 0.5 abstentions were produced on instruction, so single-side accuracy reflects prompt compliance as much as capability.
- Later rounds in the crossed branches were not regenerated after the replaced statement, so the counterfactual history is not fully self-consistent. Other players may break their own statements, but only the target is labeled.
- The game has two dishes and one-sentence commitments. Real agent interactions are far more ambiguous.
- Costs are server-reported usage. They are not invoices and exclude failed calls that returned no usage.

Dual-use considerations. The practical lesson for an actor who wants to evade *this* detector is that hiding either the public commitment or the action collapses the monitor to 0.5, which the frozen rule treats as "no violation". This is a known weakness of one-sided oversight, and the report presents it as a reason to require both sides. It is not a new concealment technique. We provide no covert signalling scheme, no private-channel construction, and no attack that changes the public record while preserving a violation. Publishing the exact prompt and threshold makes the monitor easier to audit, and it also makes surface-level gaming easier for an agent told the rule. We did not train or optimize any agent against the monitor. The crossing procedure rewrites a dish after speech to build a test set, and it must not be described as evidence that an agent chose to defect. No human subjects or personal data were involved.

Preexisting work and AI assistance. `llm-trust-breaking/` (Shi et al.) and `trace-guard/` (Nguyen et al.) predate the sprint. They serve as the game harness and as the source of the consistency question. They are not our work and should be read as preexisting. Everything in `weekend_2_2/` was written during the sprint: the four-view protocol, the 2x2 crossing, the five-round target-last process, the mixed-seat tables, the natural-play runner, the extractor and tracker, the direct-label reviewers, the statistics, and this report. An AI coding assistant was used to write and test code and to draft this report from the saved result files. The author takes responsibility for all claims. Every number in this report is copied from `REPORT_ZH.md`, `score_analysis.json`, `verification.json`, or `metrics.json` in the result directories.

Appendix B. Reproducibility and file map

Experiment	Directory	Independent games	Trajectories	Monitor calls	Cost (USD)
Positive control (rule)	<code>weekend_2_2/runs/rule_baseline</code>	10 scripted groups	40	160 (rule)	0
E1 balanced multi-round	<code>endogenous_rounds_results/role_matrix_27_20261026</code>	3	108	432	2.83
E2 mixed-model table	<code>minority_player_results</code>	6	216	864	6.05
E3 natural play	<code>naive_result</code>	6 (1 non-violation)	54	216	0.93

Experiment	Directory	Independent games	Trajectories	Monitor calls	Cost (USD)
Development check	endogenous_rounds_results (root)	1	4	16	0.20

Each result directory contains the frozen `plan.json` and `runtime_contract.json`, the source snapshot (`source_snapshot/`), all HTTP calls (`http_calls.jsonl`), the full transcripts (`GAME_TRANSCRIPT(S)_ZH.md`), per-row scores, and `verification.json`. That file confirms the shared four-round prefix, recomputed payoffs, class balance, the absence of hidden fields in monitor evidence, and extractor blindness to actions. Offline tests: 80 to 86 paired_crossing tests (depending on run version), plus 3 to 4 `endogenous_rounds` and 3 `minority_players` tests, passed before each run.

Example command (E1 development configuration):

```
python weekend_2_2/endogenous_rounds/run.py run --out weekend_2_2/endogenous_rounds_results --agents 5 --rounds 5 --games 1 --label-review-models openai/gpt-4.1 anthropic/claude-sonnet-4.5 --resume
```

Monitor-view to label mapping per trajectory in the balanced designs (round 5 only):

Target statement	Target dish	Label
"I will order CHEAP"	CHEAP	non-violation
"I will order CHEAP"	EXPENSIVE	violation
"I will order EXPENSIVE"	EXPENSIVE	non-violation
"I will order EXPENSIVE"	CHEAP	violation